DOCUMENT RESUME

ED 302 591 TM 012 661

AUTHOR Hedge, Jerry W.; Laue, Frances J.
TITLE Can Appraisers Rate Work Performance Accurately?
PUB DATE Aug 88
NOTE 23p.; Paper presented at the Annual Meeting of the American Psychological Association (96th, Atlanta, GA, August 12-16, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Information Analyses (070) -- Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Evaluation Criteria; Evaluation Methods; *Evaluation Problems; *Performance Factors; *Personnel Evaluation; Reliability; Research Problems; Validity
IDENTIFIERS *Accuracy

ABSTRACT

The ability of individuals to make accurate judgments about others is examined and literature on this subject is reviewed. A wide variety of situational factors affects the appraisal of performance. It is generally accepted that the purpose of the appraisal influences the accuracy of the appraiser. The instrumentation, or tools, available to the appraiser, also influences ratings of the appraisee. Differences among the dimensions or traits also are situational factors that affect appraisal. Much accumulated evidence about appraiser accuracy is circumstantial at best. Renewed researcher interest in the individual differences among appraisers is warranted. These factors include personality, ability, background, and vocational interest. It has also been suggested that contributing factors to appraiser accuracy include aptitude, work motivation, job satisfaction, opportunity to observe ratees, and situational constraints. Appraisers may not be as inaccurate as much previous research implies. Twenty-eight references are listed. Three tables and one figure summarize research findings. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED302591

Can Appraisers Rate Work Performance Accurately?

Jerry W. Hedge

Frances J. Laue

Universal Energy Systems, Inc.

San Antonio, TX

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

JERRY W. HEDGE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the Annual Meeting of the

American Psychological Association

Atlanta, Georgia, August 1988.

TM012661

Can Appraisers Rate Work Performance Accurately?

Job performance measurement (JPM) has been a favored topic of research and armchair theorizing for many years. Over the last several decades the JPM zeitgeist has moved from measurement instruments and appraiser training to the cognitive processes involved in appraising others. In the 1960s the Behaviorally Anchored Rating Scale (Smith & Kendall, 1963) was hailed as the cure-all for performance appraisal, but by the late 1970s researchers had begun to conclude that this method was not as superior as once believed. This led Landy and Farr (1980) to suggest a moratorium on format research. With instrumentation work on the ebb, researchers turned their attention to rater training as the "hope of the future," but early optimism has begun to decline as no approach has pushed rating accuracy levels to new heights (see Dickinson & Silverhart, 1986; Dickinson & Hedge, 1988).

The impetus for these shifts in research focus from one topical area to another has been a frustrating inability to discover the one magical method that insures high levels of appraisal accuracy. Consequently, a number of cognitive studies (e.g., DeNisi, Cafferty, & Meglino, 1984) have begun to research the processes by which appraisal information is acquired, stored, retrieved, and integrated to form an appraisal judgment. Much of this work has focused on whether people reason well or make good decisions, with inaccuracy synonymous with "shortcomings" of judgment.

At the heart of each of these topical areas is an intent to improve the accuracy with which appraisals are made. Casual reflection on much of this literature will lead the reader to conclude that appraisers are unable to make accurate ratings. However, with the exception of the recent work on cognitive processes involved in appraising others (which assumes or inadvertently suggests appraiser fallibility), appraisal research has typically concentrated on the traits/behaviors of the appraisee, the work environment, and their interaction. The resulting evaluation made by the appraiser is then "evaluated" to shed light on the accuracy/inaccuracy of the appraiser. We wish to re-direct this focus toward the traits/behaviors of the appraiser, the work environment, and their interaction in order to re-assess appraiser accuracy.

The intent of this paper is to take another look at individuals' abilities to make accurate judgments about others. We contend that appraisers may not be as inaccurate as much of the previous research implies. One premise to be explored is that certain situational factors influence the accuracy levels of ratings. A second premise of this paper is that while raters as a whole are more accurate than often believed, more research emphasis should be placed on identifying individual difference factors that lead to accurate ratings. This paper will highlight theoretical and empirical research in (a) building the case that raters do have the ability to rate accurately, and (b) suggesting that raters may be more accurate than many people believe.

Situational Factors

A wide variety of situational factors affect the performance appraisal process and as was noted earlier, research frequently focuses on how these variables impact the appraisee's performance. For example, Peters and O'Connor (1980) defined situational constraints as aspects of the immediate work situation that act, in some fashion, to interfere with the use of individual abilities and motivation in performing various jobs. Likewise, factors in the work environment can interfere with the appraiser's evaluation of the appraisee's performance. Our intent in this section is to highlight several of these studies that examine variables that influence rating accuracy, and to argue that low levels of rating accuracy need not suggest the appraiser is incapable of accurate judgments.

In order to discuss these issues, however, it is necessary to borrow a frame-of-reference proposed in the social judgment literature relevant to appraisal accuracy. In a recent review of the social judgment literature related to evaluation accuracy, Funder (1987) points out that laboratory research on "error" in social judgment has largely supplanted research that addresses accuracy issues more directly. Much interest has been focused on this body of research because of its dismal implications for the accuracy of human social reasoning. Funder stresses that these implications are illusory, because an "error" is not the same thing as a "mistake."

Drawing a distinction between the two concepts, Funder notes that an error is a judgment of an experimental stimulus that departs from a model of the judgment process. A mistake, on the other hand, is an incorrect judgment of a real-world stimulus. While errors can be highly informative about the process of judgment in general, they are not necessarily relevant to the content or accuracy of particular judgments. Errors in a laboratory may not be mistakes with respect to a broader, more realistic frame-of-reference and the processes that produce such errors might lead to correct decisions and adaptive outcomes in real life. Such a distinction can quite aptly be transferred to the performance appraisal situation.

Purpose of Appraisal

A salient example of this error/mistake distinction can be found in the performance appraisal literature on purpose of appraisal. While a substantive body of literature has not yet accumulated on this topic, it would be correct to say that I/O Psychologists have come to support the notion that purpose of the appraisal does generally influence the "accuracy" of the appraisal given by the appraiser. Studies by Bernardin, Orban, and Carlyle (1981), Zedeck and Cascio (1982), and McIntyre, Smith and Hassett (1984) all support this general claim.

When viewed from Funder's perspective, however, differential levels of accuracy of judgments obtained for appraisal feedback, personnel decisions, or research purposes must be interpreted in the broadest sense as errors rather than mistakes. When, for example, appraisal information is used as the basis for personnel decisions and the organizational norms

have dictated relatively inflated ratings for all appraisees, errors in judgment are deliberate and adaptive behaviors of the appraiser, and therefore not inaccurate. Funder's notion applied to a purpose of appraisal context, is in many ways common sensical. However, we do believe that when such studies are remembered, the appraiser's behavior is often what gets the tainted reputation as much (or moreso) than the situation that dictated such behavior. (We claim no innovative thinking, we merely wish to defend the appraiser from unwarranted abuse.)

Instrumentation

A second factor we wish to highlight that influences the appraiser's ratings of the appraisee is the instrumentation, or tools available to the appraiser. As has been routinely noted in the literature (e.g., Weitz, 1961), selection of an appraisal instrument is often governed by expediency rather than relevancy. Time and money constraints frequently influence the practitioner "charged" with selecting/developing the instrument to borrow and use; borrow, retrofit, and use; or hastily develop and use. As a result, it is not illogical to assume that the appraiser inherits a tool less than adequate for his/her needs. Once again, the onus for inaccurate ratings should not fall solely on the shoulders of the appraiser.

Given the proper instrumentation and training, appraisers can be quite reliable and accurate. A recent study illustrates this contention. Hedge, Dickinson and Bierstedt (1988) in a study of jet engine mechanic job performance used a rigorously developed task-level, go/no go measurement instrument to assess performance on steps required

to perform a variety of critical tasks in a structured work sample test. Extensive videotape-based training of test administrators preceded evaluation of jet engine mechanic performance. Data were collected from three teams of test administrators at two workshops held 2-1/2 months apart, and in a field test conducted in the interim between the two workshops. At the two workshops, videotapes of mechanic performance were shown, allowing a direct assessment of administrator accuracy. Interrater agreement was assessed at all three points in time. Table 1 shows the results of the accuracy analysis. Correlational and percent agreement accuracy indices were calculated for each test administrator team across all work sample tasks at each workshop. As can be seen, the accuracy levels for all teams were quite high at both workshops, averaging .73 for the first workshop, and .87 for the second workshop.

Insert Table 1 about here

Table 2 shows the results of interrater agreement computations via correlations and percent agreement at both workshops and the field test. High levels of interrater agreement were obtained for the three teams at all points in time. Workshop 1 correlations averaged .77, Workshop 2 correlations averaged .86, and the Field Test yielded an average correlation of .91. In summary, appraisers can be consistently accurate and reliable when the situation is structured to facilitate accurate rating behavior.

Insert Table 2 about here

Dimensional/Trait Differences

One last topic within the situational factors area warrants discussion and is closely tied to our comments on instrumentation, namely, differences among the dimensions or traits. It is our contention that differential levels of accuracy can be obtained at the dimension or trait level when rating an appraisee. Some dimensions appear inherently easier to rate than others, or said another way, the concreteness of certain factors provide more tangible or easily observable behaviors on which to base appraisals. Conceptual and empirical support for this view can be found in the personality and social psychology literature. Cronbach and Gleser (1957) described a "bandwidth/fidelity trade-off," where broad and global constructs allow the appraiser to predict a variety of behaviors at moderate levels of accuracy, whereas narrow and specific constructs allow the appraiser to evaluate at high accuracy levels within a limited range of behaviors but quite poorly outside that range.

Similarly, Hampson, John, and Goldberg (1986) noted that a variety of measures related to category breadth have been derived from the hypothesis that the cognitive availability of relevant behavioral acts varies systematically with the breadth of the trait. Allen and Ebbesen (1981) echoed this view by defining trait specificity as a dimension ranging from "specific and concrete" to "diffuse and abstract." In addition, Kirby and Gardner (1972) defined behavioral specificity as the

degree to which the trait had clear behavioral connotations, thus making it easy to picture a person behaving in the way implied by the trait. Finally, a reseach study by Funder and Dobroth (1987) reinforced these views finding that traits defining extraversion were revealed relatively directly in social behavior, and therefore easier to judge, while traits defining neuroticism were less visible and, thus, judged less accurately. An example from the performance appraisal area may help the reader to visualize the concept.

Insert Figure 1 about here

In Figure 1, taken from a BARS dimension developed by Landy and Farr (1975) numerous behavioral anchors are provided that can be readily observed by the appraiser. Still, it seems quite plausible that certain dimensions or traits are inherently more concrete. Our past experience as developers of dimensional rating forms reinforces this belief, as some dimensions generate an abundance of behavioral incidents, while other dimensions necessitate much developmental time and energy to identify the necessary anchors.

Our point in this regard is that if dimensions differ along an continuum from the abstract to the concrete, rating accuracy can be affected, and yet the blame should not arbitrarily be placed on the appraiser. Thus, it is plausible that one study using a set of concrete dimensions may find high levels of appraiser accuracy, while another study using abstract dimensions may find much lower levels of appraiser accuracy.

Two examples with some empirical data may help to reinforce this notion. Borman, Hough, and Dunnette (1976) developed a behavioral-based rating form for a manager in a problem-solving interview with a troublesome subordinate. The rating form consisted of seven dimensions: 1) controlling the interview, 2) establishing rapport, 3) reacting to stress, 4) obtaining information, 5) resolving conflict, 6) developing employees, 7) motivating employees. Hedge (1982) in a study of different training programs using the Borman et al. rating form found differential levels of accuracy by dimension both before and after training across all conditions. In addition, in almost every case significantly high levels of appraiser accuracy were obtained for the "Developing" and "Motivating Employees" dimensions, while much lower accuracy levels were obtained for the "Rapport" dimension (see Table 3).

Insert Table 3 about here

Pulakos (1984) addressed this concept of differences in accuracy across dimensions in a study of rater training effectiveness. She, too, noted that certain dimensions (i.e., "Developing the Subordinate") seem less ambiguous and are associated with potentially more concrete behaviors than other dimensions (i.e., "Establishing and Maintaining Rapport"). The Rater Accuracy Training treatment was found to be most effective on the more explicit, concrete dimensions and differential amounts of accuracy across all training methods may be related to the abstractness of the dimensions. Consequently, it may be true that

accuracy may be influenced by the dimensions used, and not as a general rule solely by the appraiser. More research needs to be designed to address this possibility.

Individual Differences Between Appraisers

Thus far the purpose of our comments and examples has been to suggest that much of the accumulated evidence concerning appraiser accuracy is circumstantial at best. The intent of the rest of this paper is to encourage renewed research interest in the capabilities of the appraiser. As was noted in the introduction, researchers have, become discouraged by the accumulating evidence that rating formats and rater training have not elevated rating accuracy to hoped for levels. Therefore our comments will focus on several appraiser individual difference variables worthy of interest.

Taft (1955) noted that in 1937, Allport suggested the ability to judge others is neither entirely general nor entirely specific, and that it would be unreasonable to expect judges of people to be uniformly successful in estimating every quality of every person. Taft also referred to a study of empathetic ability by Dymond (1949), which found that the most capable judges tended to show less variation in the accuracy with which they could judge subjects' self-ratings of six traits than did poorer judges.

Taft (1955) speaking from a social judgment perspective drew attention to the individual's ability to judge people. Reviewing relevant literature within the social psychology and personality literature, Taft identified a number of factors consistently found to be

correlated with the ability to judge the personality characteristics of others. These factors included age, high intelligence, good emotional adjustment, and social skill.

Within the performance appraisal domain few studies have researched appraiser ability. Borman (1979) using college students who had completed a series of inventories tapping various personality, ability, background, and vocational interest variables, correlated these data with rating accuracy indices of videotaped performances. The most consistently high correlates of accuracy were found to be intelligence, personal adjustment, and detail orientation, accounting for approximately 17 percent of the variance in rating accuracy. Borman concluded that rating accuracy is related to appraiser individual differences. Only a few published studies have reported studies of individual differences and rating accuracy. Variables such as cognitive complexity, mental alertness, and self-esteem were found to be unrelated to accuracy while selective attention ability (Cardy & Kehoe, 1985) and beliefs about human variability (Wexley & Youtz, 1985) were found to correlate with accuracy.

Another research study that provides some additional support to Borman's work was done by Dickinson, Hedge, and Teachout (1987). In an investigation of appraiser attributes that contribute to rating accuracy, variables such as aptitude, work motivation, job satisfaction, opportunity to observe ratees, and situational constraints were studied. General ability (as measured by a variety of aptitude tests) and situational constraints were found to be significant predictors of rating accuracy. Thus, while only meager evidence exists on appraiser accuracy, the findings are a positive first step.

Summary

Our goal in writing this paper has been to draw attention to a relatively neglected research area that may offer some hope of improving the accuracy with which appraisees are evaluated. Few studies have been conducted in the performance appraisal area on appraiser individual differences. Yet Borman's (1979) work and related information from the social psychology/personality literature suggest more research should be done in this area. In addition, we hope that our musings about the lack of substantiated findings concerning appraiser inaccuracy will at least stimulate some thoughtful introspection about these issues, and in turn, convince some researchers to empirically examine appraiser attributes as they relate to appraiser accuracy.

REFERENCES

Allen, R. B., & Ebbesen, E. B. (1981). Cognitive processes in person perception: Retrieval of personality trait and behavioral information. Journal of Experimental Social Psychology, 17, 119-141.

Allport, G. W. (1937). Personality: A psychological interpretation. New York: Holt, Rinehart & Winston.

Bernardin, H. J., Orban, J. A., & Carlyle, J. J. (1981). Performance ratings as a function of trust in appraisal, purpose for appraisal, and rater individual differences. Proceedings of the Academy of Management, 311-315.

Borman, W. C. (1979). Individual differences correlates of accuracy in evaluating others performance effectiveness. Applied Psychological Measurement, 3, 103-115.

Borman, W. C., Hough, L., & Dunnette, M. (1976). Performance ratings: An investigation of reliability, accuracy, and relationships between individual differences and rater error. Minneapolis: Personnel Decisions, Inc.

Cardy, R. L., & Kehoe, J. P. (1984). Rater selective attention ability and appraisal effectiveness: The effect of a cognitive style on the accuracy of differentiation among ratees. Journal of Applied Psychology, 69, 589-594.

Cronbach, L. J., & Gleser, G. C. (1957). Psychological tests and personnel decisions. Urbana: University of Illinois.

DeNisi, A. S., Cafferty, T. P., & Meglino, B. M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. Organizational Behavior and Human Performance, 33, 360-396.

Dickinson, T. D., & Hedge, J. W. (1988). Training to improve the accuracy and validity of performance measurements, Brooks Air Force Base, Texas: Air Force Human Resources Laboratory final report.

Dickinson, T. D., & Silverhart, T. A. (1986). Training to improve the accuracy and validity of performance ratings. Paper presented at the annual meeting of the American Psychological Association, 1986.

Dickinson, T. D., Hedge, J. W., & Teachout, M. S. (1987). Accuracy of performance ratings: A structural model of rater attributes. Paper presented at the annual meeting of the American Psychological Association, 1987.

Dymond, R. F. (1949). A scale for the measurement of empathic ability. Journal of Consulting Psychology, 13, 127-133.

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. Psychological Bulletin, 101, 75-90.

Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Properties associated with inter-judge agreement. Journal of Personality and Social Psychology, 99, 409-418.

Hampson, S. E., John, O. P., & Goldberg, L. R. (1986). Category breadth and hierarchical structure in personality: Studies of assymmetries in judgments of trait implications. Journal of Personality and Social Psychology, 51, 37-54.

Hedge, J. W. (1982). Improving the accuracy of performance evaluations: A comparison of three methods of performance appraisal training. Unpublished doctoral dissertation, Old Dominion University.

Hedge, J. W., Dickinson, T. D., and Bierstedt, S. A. (1988). The use of videotape technology to train administrators of Walk-Through Performance Testing (AFHRL-TP-87-71). Brooks Air Force Base, Texas: Air Force Human Resources Laboratory.

Kirby, D. M., & Gardner, R. C. (1972). Ethnic stereotypes: Norms on 208 words used in their assessment. Canadian Journal of Psychology, 26, 140-154.

Landy, F. J., & Farr, J. L. (1975). Police performance appraisal. Final Technical Report to the Department of Justice, Law Enforcement Assistance Administration.

Landy, F. J., & Farr, J. L. (1980). Performance rating. Psychological Bulletin. 87, 72-107.

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of ratings. Journal of Applied Psychology, 69, 147-156.

Peters, L. H., & O'Connor, E. J. (1980). Situational constraints and work outcomes: The influence of a frequently overlooked construct. Academy of Management Review, 5, 391-397.

Pulakos, E. D. (1984). A comparison of training programs: Error training and accuracy training. Journal of Applied Psychology, 69, 581-588.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47, 149-155.

Taft, R. (1955). The ability to judge people. Psychological Bulletin, 52, 1-23.

Weitz, J. (1961). Criteria for criteria. American Psychologist, 16, 228-231.

Wexley, K. N., and Youtz, M. A. (1985). Rater beliefs about others: Their effects on rating errors and rater accuracy. Journal of Occupational Psychology, 58, 265-275.

Zedeck, S., & Cascio, W. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. Journal of Applied Psychology, 67, 752-758.

Table 1

Rater Accuracy by Testing Time and Test Administrator Team

Test Administrator Team	Testing Time	
	Workshop 1	Workshop 2
Team 1	.772	.922
	(73.9)	(96.1)
Team 2	.794	.907
	(86.8)	(95.0)
Team 3	.635	.770
	(67.2)	(78.5)

Note. Primary values reported are correlations of ratings and target scores; numbers in parentheses are percent agreement values.

Table 2

Interrater Agreement by Testing Time and Test Administrator Team

	Testing Time		
Test Administrator Team	Workshop 1	Field Test	Workshop 2
Team 1	.786	.882	.918
	(74.4)	(78.7)	(83.6)
Team 2	.813	.947	.886
	(84.9)	(90.0)	(90.0)
Team 3	.712	.904	.764
	(76.4)	(85.8)	(84.1)

Note. Primary values reported are correlations; numbers in parentheses are percent agreement values.

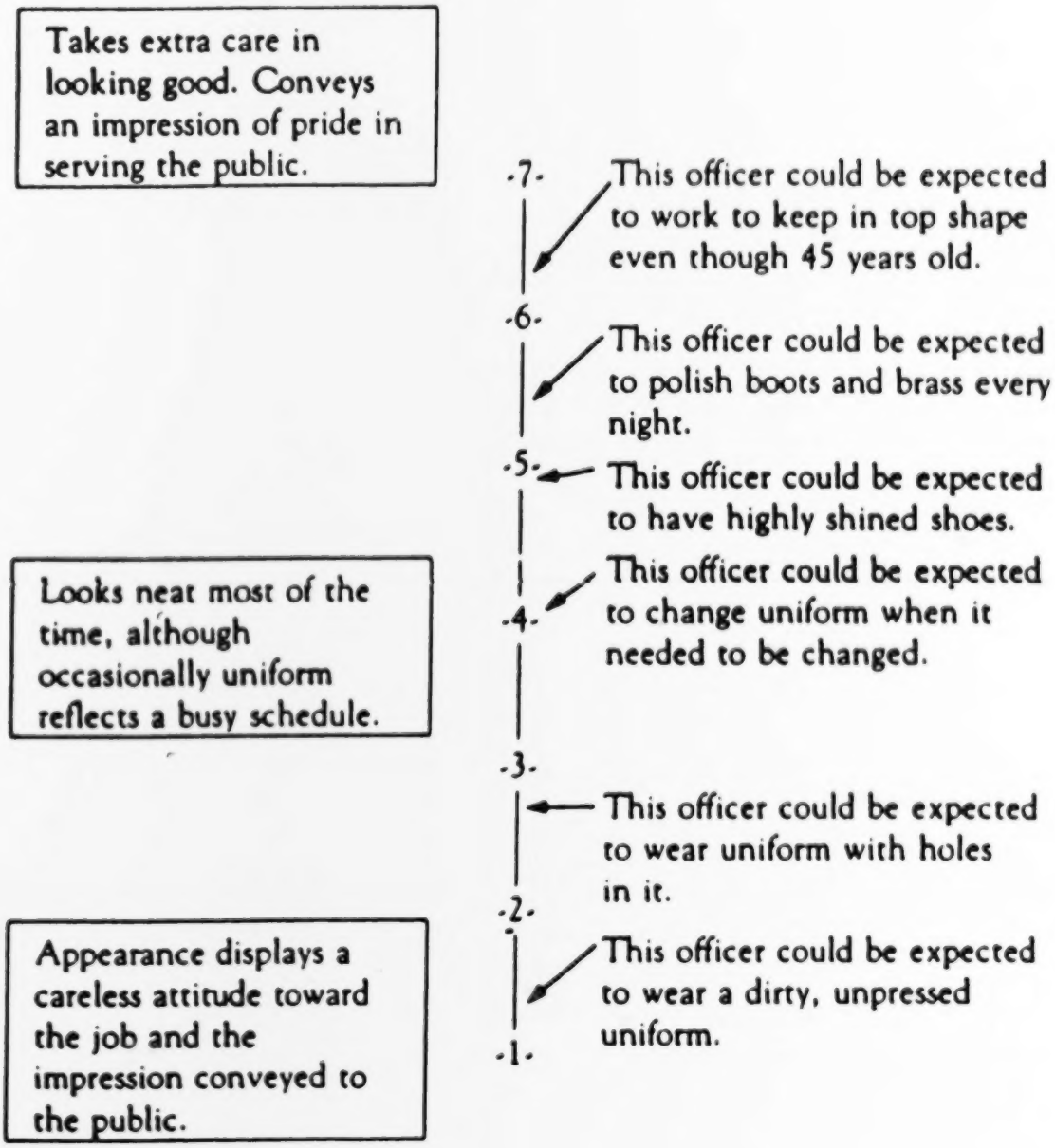
Table 3

Correlational Accuracy Values for All Dimensions

Dimension	Correlation
Motivating Employees	.73
Developing Employees	.72
Structuring/Controlling the Interview	.67
Resolving Conflict	.59
Obtaining Information	.56
Reaction to Stress	.51
Rapport	.36

Note: Correlational values are averaged across testing times and training programs.

Figure 1. BARS Dimensional Behavioral Anchors


Takes extra care in looking good. Conveys an impression of pride in serving the public.
-7- This officer could be expected to work to keep in top shape even though 45 years old.
-6- This officer could be expected to polish boots and brass every night.
-5- This officer could be expected to have highly shined shoes.
Looks neat most of the time, although occasionally uniform reflects a busy schedule.
-4- This officer could be expected to change uniform when it needed to be changed.
-3- This officer could be expected to wear uniform with holes in it.
-2-
Appearance displays a careless attitude toward the job and the impression conveyed to the public.
-1- This officer could be expected to wear a dirty, unpressed uniform.

NOTE: From Landy and Farr (1975) as in Bernardin and Beatty (1984).